The Use of Automated Critics to Improve the Fusion of Marginal Sensors for ATR and IFFN Applications *

David J. Miller and David L. Hall
The Pennsylvania State University
Department of Electrical Engineering
University Park, PA 16802-2701
(814) 865-6510, (814) 863-4155
miller@strider.ee.psu.edu, dlh28@psu.edu

Abstract

A basic goal of multi-sensor data fusion is to increase the accuracy and reliability of inferences by combining data and information from multiple sources. In particular, applications such as automatic target recognition (ATR) and identification-friend-foe-neutral (IFFN) processing seek to characterize, classify, and ultimately identify targets of interest such as aircraft, tanks, or enemy units. Ideally, the use of multi-sensor data from non-commensurate sensors (viz., sensors observing fundamentally different physical phenomena) would improve the ability to identify targets by broadening the physical baseline of observation. For example, the use of a combination of acoustic, seismic, infra-red, and radar data has the potential to improve the ability to characterize ground-based targets. In addition, the use of a broad range of physical measurements improves the ability to counter an enemy's information warfare efforts.

There are several circumstances, however, in which the fusion of multi-sensor data actually produces worse results (on average) than can be achieved by an individual sensor. That is, the fused results are less accurate and less reliable than those of the best individual contributing sensor. As one example, sensor data may be incorrectly weighted, due to a lack of knowledge of the dynamic sensor performance in realistic operating conditions. Another example most germane to the present work is the case where decisions from one or more of the contributing sensors have accuracy less than 50 percent. It is well-known that decision-level fusion schemes, such as voting techniques, produce unreliable results when the accuracy of the contributing sensors is less than 50 percent. Unfortunately, such relatively poor performance is not uncommon in applications such as IFFN and ATR. This is particularly true in anticipated information warfare conditions.

Recently, Miller and Yan [13, 12, 11] have developed a novel approach to mitigate the effects of poor sensor performance, by introducing what they refer to as an *expert-critic* technique. The concept involves the use of so-called automated *critics* to monitor individual sensor performance (or the performance of a human expert), so as to determine when a sensor or information source is likely to be reliable, and when it is likely to be faulty. Sensor data are combined using either hard or soft voting techniques modulated by the critics, which *censor* individual sensors when their performance is judged likely to be faulty. The expert-critic technique was validated both empirically as well as theoretically in [13, 12, 11] where it was shown that critic-driven voting is successful *even when the average voter accuracy is less than 50 percent.* The critics can be implemented using a variety of techniques, such as models of sensor performance, neural networks, fuzzy logic, or rule-based systems.

This paper presents an overview of the techniques developed by Miller and Yan and describes their applicability to ATR and IFFN. The results of some numerical experiments are summarized, demonstrating the performance improvements resulting from the use of critic-monitored information fusion.

*Work of the first author was supported in part by the National Science Foundation under grant no. IIS-9624870.

Form SF298 Citation Data

Report Date *("DD MON YYYY")* 00001999	**Report Type** N/A	**Dates Covered (from... to)** *("DD MON YYYY")*

Title and Subtitle The Use of Automated Critics to Improve the Fusion of Marginal Sensors for ATR and IFFN Applications	**Contract or Grant Number**
	Program Element Number
Authors Miller, David J.; Hall, David L.	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) The Pennsylvania State University Department of Electrical Engineering University Park, PA 16802-2701	**Performing Organization Number(s)**
Sponsoring/Monitoring Agency Name(s) and Address(es)	**Monitoring Agency Acronym**
	Monitoring Agency Report Number(s)
Distribution/Availability Statement Approved for public release, distribution unlimited	
Supplementary Notes	
Abstract	
Subject Terms	
Document Classification unclassified	**Classification of SF298** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** unlimited
Number of Pages 11	

1 Introduction

The fundamental concept of pooling individual decisions or opinions, to make group or collective decisions, forms the basis for reliable decisionmaking in a number of disciplines and institutions. Early (and continuing) utility for these techniques can be found in the voting assemblies of political and economic institutions. Voting techniques were early on analyzed by Laplace and Condorcet [1]. The analysis of these methods is still an active research area [1],[3]. Within the engineering discipline, voting and related combining techniques have been proposed for sensor fusion in the detection and automatic target recognition communities [8],[18],[2]. These methods have also been developed and applied in the statistical pattern recognition/neural networks/machine learning literature, where they are often denoted *ensemble classification methods* [6]. It is useful to bring out the connections between these combining techniques as studied in different domains. This is especially true here since the critic-driven paradigm was inspired by a consideration of the limitations of voting methods, was developed and applied in the context of ensemble classification[13],[11], and is here being introduced to the detection and sensor fusion communities. While the terminology, motivation, problem statement, and associated assumptions are somewhat different for the distinct problem domains, the basic combining techniques, in particular hard and soft voting, are common to these domains and are thus in some sense "universal". Likewise, it is expected that the critics approach, which extends hard and soft voting to include confidence measures on the individual voters, will also have general applicability. In the sequel, we will review the development of the critic-driven paradigm with the principal aim of exposing its applicability both for ensemble classification (as in [13, 12, 11]) and, particularly, for sensor fusion/detection. Since the method was conceived in the ensemble classification context, it is most natural to review development of the approach in this setting. However, throughout – in giving motivation, defining the problem, formulating solutions, and applying the resulting techniques – we will identify issues specific to sensor fusion as they arise. Moreover, when necessary, we will suggest how the approach may need to be modified to address particular sensor fusion objectives and/or constraints.

2 Critic-Driven Combining for Classification and Sensor Fusion

2.1 Motivation

In both multisensor fusion and ensemble classification, voting techniques are commonly employed for pooling decisions. In the multisensor fusion case, each sensor may make a local decision based on its raw sensor output. Voting methods are then used for combining the decisions. In general, combining local decisions increases accuracy, as more information is being leveraged in making the ultimate decision. Particular voting schemes may provide optimal detection under suitable conditions. In the ensemble classification case, multiple individual classifiers may be trained for a variety of reasons:

1. poor local optimum traps of the cost objective used for training (which may limit effectiveness of a single trained model);
2. insufficient training data for building a large single classifier;
3. lack of knowledge concerning the optimal classifier structure and/or the optimal feature set for discrimination. This suggests pooling classifiers with heterogeneous structures (decision trees, neural networks, rule-based classifiers, etc.) and/or that use different feature spaces.

Each of these factors may contribute to ensemble methods outperforming standard, stand-alone classification even if the classifiers that comprise the ensemble all use the same feature observations.

The effectiveness of combining can be given some simple analytical justification for majority-based voting, assuming the voters all make errors or correct decisions independent of each other. Under this assumption the number of correct voters is given by the binomial distribution. If the individual classifier error rate is $p < 0.5$, it is a known result (Condorcet's theorem) [1],[3] that for odd (or even) number of voters N, the correct decision rate for the voting system increases with increasing N, going to one as $N \rightarrow \infty$[1]. Alternatively, if $p > 0.5$, the correct decision rate *decreases* with increasing N. Although this analysis is based on an independence assumption which

[1]This result is attributed to the Marquis de Condorcet (18th century) [1],[3], who addressed the problem in a jury context.

does not generally hold in practice, it does provide some analytical justification for using voting techniques, *i.e.*, it reasonably indicates that performance should improve with increasing number of voters. However, it also identifies a fundamental requirement on the accuracy of individual voters for the success of the scheme. While it is often reasonable to assume $p < 0.5$, there are a number of ensemble classification and distributed detection situations where this condition may in fact be violated:

1. when there are numerous classes to discriminate or detection hypotheses to evaluate;
2. in the classification setting, when there is insufficient training data for building the individual classifiers; similarly, in the sensor fusion setting, when the local sensor distributions are inaccurately known;

   when each individual sensor provides only weak discriminatory information;
4. in non-stationary settings;
5. in a multistage classification/detection setting, where the system is used as a "final arbiter" stage to classify the difficult samples rejected by (simpler) previous stages.

The fundamental limitation $p < 0.5$ inspired the critic-driven paradigm of Miller and Yan, which was developed specifically to overcome this requirement. It should be mentioned that, in the ensemble classification case, there are methods which can overcome $p > 0.5$ by jointly optimizing a feature-dependent combining rule given common training data. In particular, we note methods such as mixture of experts [10] and related approaches e.g. [17] which may allow effective combination of even (extremely) weak classifiers. However, the required joint optimization is often impractical. Moreover, there are increasingly important scenarios involving, e.g., distributed databases where common, commensurable training data is *not* available, and hence where optimization of the combining rule is not possible. Such feature-dependent combining rules are also generally not applicable in distributed detection, especially for the limited bandwidth case where the local sensors communicate their decisions (rather than the raw sensed data) to the central decision function. For this detection case, $p < 0.5$ is typically considered to be a rigid requirement.

2.2 The Critic-driven Paradigm

The abovementioned circumstances necessitate combining methods that pool decisions rather than raw features. For ensemble classification, the implication will then be that common, commensurate training data will not be required; nor will a complex optimization procedure be required for learning an input-dependent combining rule. In the distributed detection case, the implication is that sensors can communicate low-bandwidth decisions, rather than (more informative but costly) raw features, to the central combiner. While decision-level combining yields these practical advantages, this practicality is achieved by sacrificing performance – as mentioned before, simple combining methods such as standard voting techniques are fundamentally limited in that they are *only* successful when the individual voter's error rate p is less than 0.5. Miller and Yan sought to overcome this restriction by developing a new decision-level combining scheme. The difference between their work and previous voting techniques was their introduction and judicious use of a *critic*, specific to each voter, which evaluates the voter's decisions. Given access to the same input feature vector/raw sensor signal used by the voter, the critic tries to predict whether the voter's decisions are *valid* or *bogus*. Critic-driven combining schemes incorporate both the collection of voter decisions and the associated critic evaluations in forming an ultimate decision. In the ensemble classification setting, each critic is trained after its voter/expert, on the expert's training set, with supervision information indicating whether or not the expert's decisions agree with the true class labels – essentially, the critic is trained to be an expert *on* the expert. In the distributed detection problem, critics could be derived from models for the inaccuracy of the assumed sensor distributions (e.g. under different operating conditions), from separate information sources, or even from human intervention. The potential advantage of the critic-expert methodology is best seen for the case where there are many classes/detection hypotheses. Since the expert tackles a multiclass (perhaps $\gg 2$ class) problem and the critic only needs to solve a two-class problem (even though a potentially difficult one), the critic's estimates should be more accurate than the expert's on average. In the ensemble classification setting this has been confirmed experimentally [13, 11]. The suggestion is then to incorporate the critic's opinion within the rule of combination, so as to achieve more reliable decisions. Several different combining rules were proposed in [13].

This use of a critic for ensemble classification and distributed detection is somewhat related to [19], where local accuracy estimates were used to determine which classifier makes the decision for a given datum. The concept of applying measures of reliability or confidence to expert decisions is also sometimes used for refereeing conferences and journals, where one is asked to provide both a numerical evaluation of paper quality and a measure of confidence. We also note methods such as [7] which, similar to the critics approach, use a paradigm wherein two-class decisions contribute to the formation of multiclass decisions. However, [7] does not use the notion of a critic and requires joint optimization of its simple classifiers based on a common training set.

Formulation

Both hard voting and soft averaging critic-driven schemes were proposed in [13]. Both types were proposed specifically for ensemble classification and may require minor modifications for the detection domain. In particular, the hard voting techniques assume that all classification errors have equal importance, whereas in the detection case there may be prescribed target values for the false alarm or the false reject probabilities. The soft averaging methods assume that *a posteriori* probabilities from each expert-critic pair are aggregated by the central combiner. However, in the constrained bandwidth detection setting these probabilities may need to be quantized prior to their transmission to the central combiner. These issues will be further addressed as the techniques are next discussed.

Critic-driven Hard Voting

Several critic-based extensions of standard voting methods were proposed in [13]:

Critic-driven majority-based voting: In this case, each expert votes for the class which it predicts if the critic assesses that the expert prediction is valid. Otherwise, the expert *abstains* from voting. The total number of votes (K) equals N minus the number of abstentions. All votes are given equal weight. If any class receives $l > \frac{K}{2}$ votes, then that class is the one predicted by the ensemble. Otherwise the datum is rejected. In order to place emphasis on minimizing errors of a certain type, associated with a particular class (as may be required for the detection problem), one could instead use an unequal weighting of the votes from the different classes.

2. *Modified critic-driven majority voting*: An alternative rule is motivated by the following fact: the probability of correct decision for majority-based voting with N voters, N even, is *less than* the probability of correct decision with $N-1$ voters. This suggests a modification of the critic-based scheme just described wherein, when K is even, one voter is *dropped* prior to vote tallying. In practice, the voter with least confidence from its critic could be dropped. This modified scheme has two significant advantages over the former one. First, as one might expect, this scheme achieves a greater correct decision rate. Second, it turns out that this scheme greatly simplifies performance analysis.

3. *Critic-driven plurality voting*: This scheme is similar to a), except that now the class receiving the most votes is predicted by the ensemble. Rejections thus only occur when there is a tie.

The critic-driven voting schemes clearly provide potential for improving the reliability/"efficiency" of the voting ensemble, by removing unreliable voters. A second advantage for distributed detection is that, by censoring sensor transmissions, one can further reduce the transmission bandwidth.

2.3.2 Critic-driven Averaging of Probabilities

In the case of soft averaging, there are several distinctions that must be made between the ensemble classification and distributed detection problems. First, in the ensemble classification case, each classifier can in principle use the same set of feature observations. However, in distributed detection, each detector only has access to its locally sensed observations. This distinction mainly affects our choice of mathematical notation, without seriously impacting the formulated combining rules. A second difference is the fact that, in the detection problem, the local sensor information needs to be transmitted, possibly with crude precision so as to reduce bandwidth. In the sequel we will consider the formulation within the context of ensemble classification, where no

such communication need will be assumed. Thus, the central combiner will aggregate "perfect" information in the form of *a posteriori* probability estimates produced by each classifier. For distributed detection, we emphasize that these probabilities would need to be quantized.

Consider classification of a feature vector $\underline{x} \in \mathcal{R}^d$ into one of C classes. Assume there are N voters/experts. We assume each expert produces estimates of the *a posteriori* class probabilities, i.e. $P_e^{(j)}[k|\underline{x}]$, $k = 1, \ldots, C$ $j = 1, \ldots, N$, with "e" denoting expert. Each critic also produces probabilities $P_c^{(j)}[b|\underline{x}]$, where $b \in \{0, 1\}$, with "1" indicating a *valid* assessment and "0" a *bogus* assessment. Here "c" denotes the critic. In [13, 11], methods for combining these probabilities were derived based on information-theoretic principles. Here we summarize that development.

A loosely stated objective for the combiner is to "agree with" expert probabilities, to the extent that they are valid, as estimated by the critic. Information theory suggests in this case use of a cross entropy (Kullback-Leibler distance) criterion – a measure of dissimilarity between probability mass functions that has been given axiomatic justification [15]. However, since cross entropy is an asymmetric cost, there are two possible objectives.
The Geometric Average Rule: If we view the expert probabilities as *priors*, then we will choose the combined probabilities $\{P[k|\underline{x}]\}$ as the posteriors minimizing the average cross entropy cost:

$$\sum_{j=1}^{N} w_j(\underline{x}) D(\{P[k|\underline{x}]\} || \{P_e^{(j)}[k|\underline{x}]\}). \tag{1}$$

Here, $D(\{P[k|\underline{x}]\} || \{P_e^{(j)}[k|\underline{x}]\}) \equiv \sum_{k=1}^{C} P[k|\underline{x}] \log\left(\frac{P[k|\underline{x}]}{P_e^{(j)}[k|\underline{x}]}\right)$, the standard cross entropy cost between pmfs, with the weighting function $w_j(\underline{x}) = \frac{P_c^{(j)}[1|\underline{x}]}{\sum_{l=1}^{N} P_c^{(l)}[1|\underline{x}]}$, a probabilistic measure of the critic's confidence in its expert[2]. After minimizing (1) over $\{P[k|\underline{x}]\}$ subject to constraints ensuring a pmf solution, we obtain the "geometric average" estimates:

$$P[k|\underline{x}] = \frac{\prod_{j=1}^{N} (P_e^{(j)}[k|\underline{x}])^{w_j(\underline{x})}}{\sum_{m=1}^{C} \prod_{j=1}^{N} (P_e^{(j)}[m|\underline{x}])^{w_j(\underline{x})}} \quad k = ,\ldots, C. \tag{2}$$

The chosen class is then the one with maximum *a posteriori* probability.
The Arithmetic Average Rule: Alternatively, we can interpret the experts as *posterior* probabilities and seek a *prior* probability agreeing with each posterior, to the extent that it is valid. In this case, we obtain the "arithmetic average" rule:

$$P[k|\underline{x}] = \sum_{j=1}^{N} w_j(\underline{x}) P_e^{(j)}[k|\underline{x}]. \tag{3}$$

Equation (3) is a generalization of simple averaging [4], and specializes to it with the choice $w_j(\underline{x}) = \frac{1}{N}$. Both (2) and (3) are effective schemes, with neither dominating the other in all cases. In particular, we have found that while (2) often achieves better results than (3), it may produce less reliable results when some experts give probabilities close to zero.
An Improved Rule: While both (2) and (3) outperform simple averaging [13, 11], neither approach gleans all the information contained in the ensemble. In particular, if we consider the case where $P_c^{(j)}[1|\underline{x}] = 0$, we see that in both (2) and (3), expert j effectively abstains from contributing its estimates. However, a zero probability from a critic is actually quite informative – it reasonably indicates that the expert's predicted ("winning") class should be excluded. This suggests the following approach: conditioned on critic j's validation of its expert, the pmf $\tilde{P}_e^{(j)}[k|\underline{x}, b_j = 1] = P_e^{(j)}[k|\underline{x}]$ is posited; conditioned on the critic's rejection, a uniform pmf is posited over all classes excluding the expert's predicted winner:

$$\tilde{P}_e^{(j)}[k|\underline{x}, b_j = 0] = \begin{cases} \frac{1}{C-1} & \text{if } k \neq c^* \\ 0 & k = c^*, \end{cases} \tag{4}$$

[2]Other choices for the weights $w_j(\underline{x})$ that are monotonically increasing in $P_c^{(j)}[1|\underline{x}]$ have also been found to be effective.

where $c^* = \arg\max_c P_e^{(j)}[c|\underline{x}]$. Now, the average cross entropy cost sums over $2N$ terms and the resulting estimator, assuming experts as posteriors, is:

$$P[k|\underline{x}] = \sum_{j=1}^{N} \sum_{l=0}^{1} w_{jl}(\underline{x}) \tilde{P}_e^{(j)}[k|\underline{x}, l] \tag{5}$$

2.4 Analysis of Voting Methods

One might reasonably be skeptical that the critics paradigm will offer any advantage/capability over standard voting and averaging, especially in the ensemble classification case. In particular, here the critics approach does *not* in general make use of *any* additional information. It is simply a special choice of structure and learning, with auxiliary networks trained to provide reliability on their experts. However, in [12, 11], the critic-driven schemes, and in particular modified critic-driven hard voting, were validated by the following theorem which demonstrates that critic-driven voting fundamentally extends achievable voting performance.

Theorem: Assume experts make independent errors with common rate p. Further, assume that critics make errors at common rate q independent of their experts and independent of other experts and critics. Then, the modified critic-driven hard voting correct decision rate increases with increasing integer N if $p + q < 1$.

Proof : see [12, 11].

Figure 1 (from [11]) plots the predicted correct decision rates for majority-based voting and for the modified critic-driven voting versus N for the choices $p = 0.52$ and $q = 0.45$ (realistic since they have been experimentally observed). Note that for increasing N, the critic-based performance improves while the standard voting performance deteriorates. Moreover, note that the critic-based curve is a smooth one, increasing for an increasing sequence of *integers*, consistent with the proof, while the standard majority curve is jagged. Essentially, the fundamental limitation $p > 0.5$ is overcome by forcing abstentions so as to reduce the individual *voter's* error rate below 0.5, *even when the expert rate p is above 0.5.* Thus, even if $p > 0.5$, the ensemble performance will improve with increasing number of experts/sensors so long as q can be made sufficiently small. Since q is the error rate associated with a two-class problem, it is quite possible that sufficiently accurate critics can be constructed for this purpose – experimental results have been obtained to this effect. In summary, there is a potentially much less stringent condition for successful voting in the critic-driven case than in the standard majority-based case, under an independence assumption (as will be further confirmed experimentally). Moreover, even if $p < 0.5$, the use of critics may increase the overall "efficiency" of the voting, leading to improved performance [11].

3 Experimental Results

In this section we summarize several ensemble classification experiments previously reported in [13]. There, the various combination schemes were evaluated using radial basis functions (RBFs) [14] and decision trees [5] as the basic classifier structures. These structures were used to form both the experts and the critics. For a particular training/test split, performance curves were generated for several methods to demonstrate how the conventional approaches and the critic-driven ones fare for particular choices of p and, in the critic-based case, q.

In Figure 2, simple averaging and the critic-driven "arithmetic averaging" were compared for RBF-based classification on the UC Irvine repository data set, *glass*. The 214 sample data set was equally split into training and test sets. For simple averaging, experts were designed with 16 RBF components. For the critic-based scheme, (expert,critic) pairs were designed with (16,20) RBF components. The critic networks are actually less complex than the experts, since they only have two outputs, one per class. For this experiment, $p \simeq 0.51$ and $q \simeq 0.47$. Thus, $p > 0.5$ and $p+q < 1$. Here, the theorem of section 2 (extrapolated to the soft averaging case) is essentially validated. The trend for critic-based performance is a decrease in the error rate for increasing N, while there is no improvement (and some degradation) with increasing N for standard probability averaging. Moreover, the benefit of critic-based combining over averaging is substantial.

As a second example, consider hard plurality voting on Deterding's *vowel* set. The 990 samples in this set were split into 525 training and 465 test. In this case, decision tree classifiers were used with both experts and critics consisting of 47 nodes. For this difficult example $p = 0.66$ and $q = 0.46$, i.e. $p > 0.5$ and $p + q > 1$.

Therefore, based on the analysis of section 2, we expect that the performance of both methods will degrade with increasing N. However, we see in Figure 3 that the standard voting error rate[3] stays roughly constant (even increasing a little) with increasing N, while the critic-driven rate *decreases* significantly. Moreover, critic-driven combining achieves a substantial performance advantage for increasing N. These results, which (in this case) prove the independent analysis pessimistic, can be explained from the standpoint of expert dependence. For the critic-based scheme, even though $p + q > 1$, there must be regions of significant probability mass in the feature space where $p + q < 1$, and where the experts are roughly independent (thus allowing correct decision rate improvement for increasing N). Further, it may be the case that where $p+q > 1$, the experts are dependent – thus, the correct decision rate will not necessarily decrease for increasing N, even in regions where $p + q > 1$. A corresponding argument can be applied to explain the standard voting performance.

A third example of $p > 0.5$, for majority voting on the *yeast* data set, is shown in Figure 4. The training/test split for this set was 742/742. CART classifiers and critics were designed with 31 and 47 nodes, respectively. Here $p = 0.53$ and $q = 0.41$. Again the critic-driven trend is a decreasing error rate. Also, the performance is significantly better than the standard majority curve.

4 Application to ATR and IFFN

Automatic Target Recognition (ATR) and Identification-Friend-Foe-Neutral (IFFN) are especially challenging multi-sensor applications. ATR involves use of multi-spectral data to identify (generally ground-based) targets for smart weapons, for tactical situation assessment systems, and for automated battlefield damage assessment [18]. Examples of ATR include automatic mine detection, use of ground-based sensors for target identification and tracking [16], and detection and characterization of undersea targets. Because of the extended physical baseline observed by spectrally diverse sensors, an opportunity exists for improved target characterization and classification. For example, identification of a ground vehicle such as a tank, can be improved by utilizing a combination of visible imagery (for target size and shape), infrared signature data (to characterize the engine heat), acoustic information (to characterize engine sounds), and seismic data (to establish information about target weight). This information increases the dimensionality of the observation space and improves the ability to distinguish among different types of targets [18]. The additional information is particularly important in an increasing era of information warfare [18] in which stealth, camouflage, and other techniques are utilized to deliberately confound target recognition.

ATR applications often require the fusion of non-commensurate sensor data - that is, data from sensors measuring fundamentally different physical phenomena. In order to fuse non-commensurate sensor data, feature-level or decision-level fusion techniques must be used (since accurate target models are generally not available which link observed parameters from one physical domain to another). Decision-level fusion techniques include: Bayesian inference, Dempster-Shafer methods, and voting methods [8]. In the ATR literature, Bayesian and Dempster-Shafer methods have tended to dominate over the use of voting methods. While probabilistic techniques are mathematically appealing, it can be shown that these methods yield erroneous results if a priori information about dynamic sensor performance is not available (or is incorrectly estimated). Unfortunately, dynamic sensor performance in realistic tactical environments varies widely, and is generally not modeled. For example, the performance of acoustic sensors can vary by a factor of 100, depending upon atmospheric conditions. In addition, the performance of sensors such as radar and infrared detectors can also vary widely based upon the environment and the efforts of an opponent to affect the observing environment. Hence, probabilistic techniques may be less robust in practice than voting methods, especially if the expert-critic approach is used.

The expert-critic approach described in this paper will be easily applicable to ATR. It is envisioned that an expert-critic could utilize environmental information (e.g., local atmospheric conditions, etc.), and human input (e.g., the likelihood of information warfare, information about doctrine or rules of engagement, etc.) to determine the potential accuracy of voting sensors. Information from diverse sensors can be combined to improve ATR accuracy without the usual corruption problems associated with the wide variability of sensor performance. Even simple explicit rules of thumb (e.g., acoustic sensors are good at night, but not-good at mid-day) could be encoded in an expert-critic. These explicit rules could be adapted to specific locations, tactical situations, and even specific targets of interest. In the architecture described in [16], such information could be remotely

[3]For hard majority and plurality voting, errors include rejections in the experiments.

down-loaded to tactical unattended ground sensing systems.

The comments about ATR are also applicable to IFFN applications. A challenge for IFFN is the wide variability of sensor performance as a function of engagement range and aspect angle, environmental conditions, information warfare, and other effects [9]. In addition, the proliferation of weapon systems, the agility of enemy sensor systems, and the availability of systems to corrupt the electromagnetic environment, cause IFFN to be difficult. It is not uncommon for IFFN sensors to perform with a probability of correct identification of less than 0.5. As in the ATR case, sensor performance models either do not exist, or are computationally prohibitive for tactical on-board computing. The expert-critic approach is anticipated to be very useful to improve IFFN. Again, simple rules-of-thumb could be encoded to perform the critic function.

5 Conclusions

In this work we have reviewed the critic-driven combining schemes of Miller and Yan [13, 12, 11], with an aim to exposing the applicability of these methods for ensemble classification and, in particular, for sensor fusion. A coming paper [11] extends the work summarized here in several directions. First, a novel combination rule is developed for the case where critics are "weak", i.e., where critics do not make explicit use of input features/raw sensor data. Second, while the analysis based on independence does provide insight, the inaccuracy of the independence assumption motivated development of an alternative analysis technique for predicting ensemble performance which incorporates prior knowledge on classifier/sensor dependence. This technique is based on maximum entropy statistical inference. The resulting predictions are more accurate than those assuming independence [12, 11].

References

[1] S. Berg. Condorcet's jury theorem, dependency among jurors. *Social Choice and Welfare*, 10:87–95, 1993.

[2] R. S. Blum, S. A. Kassam, and H. V. Poor. Distributed detection with multiple sensors. *Proc. of the IEEE*, 85(1):64–79, 1997.

[3] P. J. Boland. Majority systems and the Condorcet jury theorem. *The Statistician*, 38:181–189, 1989.

[4] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[5] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, 1984.

[6] T. G. Dietterich. Machine-learning research: four current directions. *AI Magazine*, pages 97–136, Winter 1997.

[7] T. G. Dietterich and E. B. Kong. Error-correcting output coding corrects bias and variance. In *Proc. of the Twelfth Intl. Conf. on Machine Learning*, pages 313–321, 1995.

[8] D. L. Hall. *Mathematical Techniques in Multisensor Data Fusion.* Artech House, Inc., Norwood, MA., 1992.

[9] D. L. Hall, P. Lapsa, and C. Voos. Improving the performance of ESM using a hierarchical data fusion approach. In *Proc. Combat Identification Systems Conference*, 1995.

[10] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures-of-experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994.

[11] D. J. Miller and L. Yan. Critic-driven ensemble classification. (To appear in IEEE Transactions on Signal Processing, 1999.).

[12] D. J. Miller and L. Yan. Some analytical results on critic-driven ensemble classification. (To appear in proceedings of the IEEE Workshop on Neural Networks for Signal Processing , 1999.).

[13] D. J. Miller and L. Yan. Ensemble classification by critic-driven combining. In *Proceedings of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 1999.

[14] J. Moody and C. J. Darken. Fast learning in locally-tuned processing units. *Neural Computation*, 1:281–294 1989.

[15] J. E. Shore and R. W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. on Info. Theory*, 26:26–37, 1980.

[16] D. C. Swanson and D. L. Hall. Real-time data fusion processing of internetted acoustic sensors for tactical applications. In *Proc. IEEE Intl. Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 443–446, 1994.

[17] N. Ueda and R. Nakano. Combining discriminant-based classifiers using the minimum classification error discriminant. In *Neural Networks for Signal Processing*, pages 365–374, 1997.

[18] E. Waltz and J. Llinas. *Multisensor Data Fusion.* Artech House, Inc., Norwood, MA., 1990.

[19] K. Woods, W. P. Kegelmeyer Jr., and K. Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE Trans. on Patt. Anal. and Machine Intell.*, 19(4):405–410, 1997.

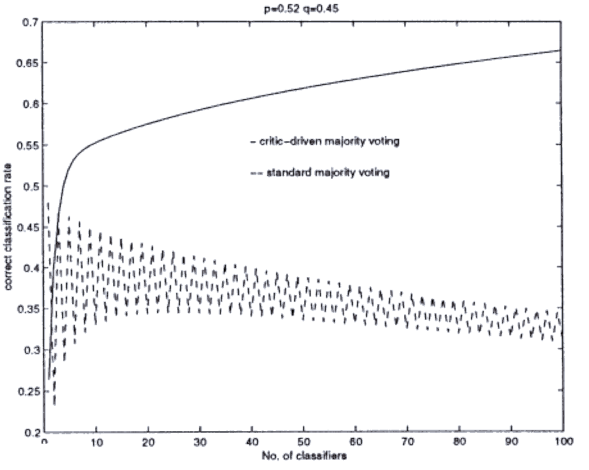

Figure 1: Analytical correct classification rates of critic-driven majority voting and standard majority voting based on an independence assumption with p=0.52 and q=0.45.

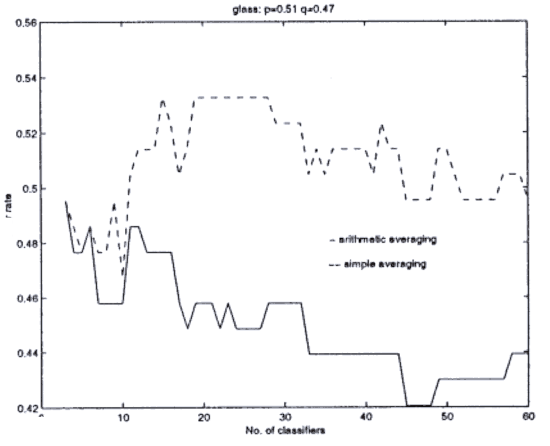

Figure 2: Error rates of RBF-based critic-driven arithmetic averaging and simple averaging for a single split of the *glass* data set.

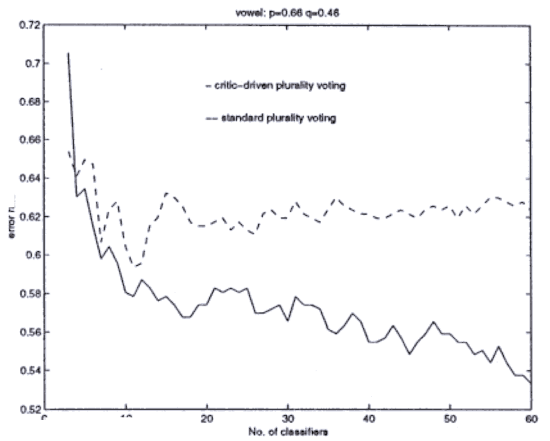

Figure 3: Error-plus-rejection rates of CART-based critic-driven plurality voting and standard plurality voting for a single split of the *vowel* data set.

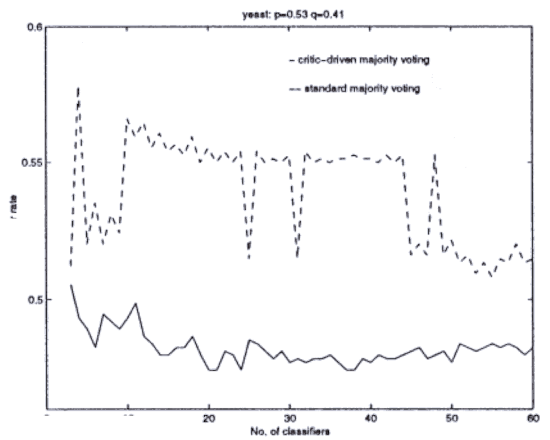

Figure 4: Error-plus-rejection rates of CART-based critic-driven majority voting and standard majority voting for a single split of the *yeast* data set.